UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Sánchez Catena, Iker. *Governança i Visualització de Dades en l'Obra Pública.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Enginyeria de Dades)

This version is available at https://ddd.uab.cat/record/337325

under the terms of the  license.

Governança i Visualització de Dades en l'Obra Pública

Iker Sánchez Catena

Resum—El present Treball de Final de Grau aborda la problemàtica de gestió i fragmentació d'informació en l'àmbit de l'obra pública dins de l'Àrea Metropolitana de Barcelona (AMB). Actualment, la manca d'una estructura de dades centralitzada i la coexistència de sistemes heretats han generat duplicitats i inconsistències que dificulten la presa de decisions. Per resoldre aquesta situació i garantir la integritat del sistema, el projecte defineix i implementa un model de governança de dades propi i una migració integral des dels sistemes heretats de l'empresa cap a un nou gestor corporatiu unificat, procés durant el qual s'han eliminat les duplicitats històriques. Paral·lelament, s'han desenvolupat processos d'Extracció, Transformació i Càrrega (ETL) encarregats de nodrir una nova base de dades complementària interna. Aquests processos integren controls automàtics per verificar contínuament la integritat i la qualitat de les dades. Aquest nou ecosistema culmina en un sistema de visualitzacions unificades que integra totes les fonts per optimitzar la presa de decisions dels gestors públics.

Paraules clau—Migració de dades, Governança de dades, ETL, Qualitat de dades, Data leakage, Aprenentatge automàtic, Obra pública.

Abstract—This Final Degree Project addresses the problem of information management and fragmentation in the field of public works within the Barcelona Metropolitan Area (AMB). Currently, the lack of a centralised data structure and the coexistence of legacy systems have generated duplications and inconsistencies that hinder decision-making. To resolve this situation and ensure the integrity of the system, the project defines and implements a custom data governance model and a comprehensive migration from the company's legacy systems to a new unified corporate manager, a process during which historical duplications have been eliminated. In parallel, Extract, Transform and Load (ETL) processes have been developed to feed a new complementary internal database. These processes incorporate automatic controls to continuously verify data integrity and quality. This new ecosystem culminates in a unified visualisation system that integrates all sources to optimise decision-making for public managers.

Index Terms—Data migration, Data governance, ETL, Data quality, Data leakage, Machine learning, Public works.

1 Introducció

El present treball es desenvolupa en el context de l'Àrea Metropolitana de Barcelona (AMB), administració pública encarregada de la planificació i gestió del territori metropolità de Barcelona, que dona servei a 36 municipis i més de 3,3 milions de ciutadans. Concretament, se centra en el Departament d'Espai Públic, format per aproximadament 160 treballadors i responsable de la promoció, disseny i execució d'obres públiques d'alt impacte ciutadà, com ara parcs urbans, infraestructures viàries, equipaments municipals i gestió del litoral.

L'entorn de l'obra pública en aquest departament genera un gran volum de dades sobre projectes i pressupostos que actualment presenten problemes severs de duplicitat i integritat. Aquestes dades es troben distribuïdes entre un sistema heretat basat en Oracle [1], fitxers Excel emmagatzemats a SharePoint [2] i serveis ArcGIS [3], fet que dificulta enormement la seva explotació i la presa de decisions. Per resoldre aquesta situació, l'empresa ha iniciat una migració cap a un nou gestor corporatiu, més modern i estructurat, que coexisteix amb una base de dades complementària interna destinada a absorbir la informació més variable i canviant (documents, imputació d'hores, plans de contractació).

Aquesta dualitat arquitectònica i la fragmentació històrica de les fonts plantegen un repte clau: evitar reproduir el descontrol d'informació del passat, que es traduïa en duplicitats entre sistemes, inconsistències en les dades d'una mateixa actuació i una elaboració d'informes lenta i manual. El projecte es justifica per la necessitat d'establir un model propi de governança de dades que decideixi quina informació es migra, quina es manté com a complementària i quina és innecessària, així com de construir l'ecosistema tècnic que en garanteixi la qualitat de manera contínua.

Per tot això, el principal objectiu del treball és el disseny, implementació i test d'un sistema de governança. Els objectius més concrets del treball són els següents:

- Realitzar una auditoria i definir un model de governança de dades.
- Dissenyar el model de la nova base de dades complementària.
- Desenvolupar els procediments de migració des del sistema heretat.
- Implementar processos ETL per a l'actualització periòdica.
- Alimentar el Geoportal [4], l'aplicatiu web intern de gestió espacial de l'AMB.
- Desenvolupar un model predictiu de compliment de subfases.
- Validar la integritat i qualitat de les dades resultants.

El desenvolupament del projecte no ha seguit una seqüència lineal estricta, sinó que s'ha adaptat a les dependències reals de cada bloc de treball. S'ha partit d'una fase

inicial d'auditoria i definició del model de governança, a partir de la qual es va iniciar la migració del sistema heretat. No obstant això, aquesta fase va quedar condicionada per dependències externes —concretament, els retards del proveïdor encarregat del nou gestor corporatiu—, fet que va obligar a aparcar-la temporalment. Aprofitant que el *pipeline* ETL i la base de dades complementària són independents del gestor corporatiu, es va prioritzar el seu desenvolupament, ja que aporta valor immediat a l'explotació de dades del Geoportal sense dependre de la disponibilitat del sistema extern. Un cop desbloquejada, es va reprendre la migració, i finalment es va abordar la capa de visualització. Aquest enfocament, basat a reordenar les tasques segons les seves dependències i el valor que aportaven, va permetre avançar de manera contínua malgrat els bloqueigs externs.

Pel que fa a les eines emprades, el projecte s'ha recolzat en un conjunt de tecnologies d'enginyeria de dades. La capa de persistència utilitza Oracle com a sistema d'origen i MySQL [5] i PostgreSQL [6] com a bases de dades de destí. Els processos de migració i ETL s'han desenvolupat en Python, fent ús de llibreries com Pandas [7] per a la manipulació de dades, s'ha utilitzat SQLAlchemy i psycopg2 per a la connexió amb les bases de dades, i biblioteques específiques per a l'accés a les fonts d'origen (Office365-REST-Python-Client [8] per a SharePoint i l'API de Python d'ArcGIS [9]). El model predictiu s'ha entrenat amb LightGBM. Per a la generació automàtica d'informes s'han emprat les llibreries python-docx, openpyxl i python-pptx, i Playwright per a la compilació dels informes a format PDF. El control de versions del codi s'ha gestionat amb Git [10] i GitHub.

La resta del document s'organitza de la manera següent. La Secció 2 descriu el model de governança i l'arquitectura de dades dissenyada. La Secció 3 detalla el procés de migració des del sistema heretat. La Secció 4 presenta el *pipeline* ETL d'integració contínua. La Secció 5 exposa la capa d'explotació de dades, amb els informes automatitzats i el model predictiu de compliment de subfases. La Secció 6 recull l'estratègia de validació aplicada i els resultats obtinguts. Finalment, la Secció 7 exposa les conclusions i les línies de continuïtat del treball.

2 Governança i Arquitectura

Aquesta secció presenta el model de governança de dades definit per al projecte i l'arquitectura resultant. En primer lloc, s'analitza l'estat del sistema heretat i les deficiències detectades durant l'auditoria inicial. A continuació, es descriu el model de governança dissenyat i el disseny de la base de dades complementària que se'n deriva, incloent-hi les regles de correspondència entre el sistema antic i el nou, les capes de tractament aplicades sobre la informació d'origen i el tractament específic de les dades de caràcter personal.

2.1 Estat del sistema heretat

L'ecosistema de dades existent està format per dues bases de dades principals que coexisteixen de forma fragmentada. La primera, originària de l'any 2015 i suportada per un motor Oracle (DB_PIO), centralitza la informació operativa dels equips —gestió d'obres i control econòmic— amb unes 76 taules. L'auditoria inicial va revelar deficiències severes en la normalització de la base de dades. Concretament, s'hi han identificat antipatrons de disseny recurrents, com relacions directes amb taules creades de forma seqüencial (un patró repetit en una vintena d'estructures similars), fet que penalitza greument l'eficiència de les consultes i n'incrementa la complexitat de manteniment. Addicionalment, s'hi han detectat un volum significatiu de camps sense ús real, relacions desnormalitzades acumulades al llarg dels anys i registres obsolets o incomplets que fan imprescindible un procés exhaustiu de *data cleansing* previ a qualsevol migració.

La segona base de dades (DB_AUX), creada l'any 2022 sobre PostgreSQL i amb 61 taules, va néixer per pal·liar les limitacions de l'esquema anterior. El seu contingut s'estructura en tres blocs. El primer bloc correspon a les dades espacials GIS (*Geographic Information System*) vinculades a l'aplicatiu web Geoportal. El segon bloc conté dades replicades i desnormalitzades del sistema de 2015, redundància introduïda històricament per adaptar formats. El tercer bloc integra dades provinents de fonts externes com SharePoint, directoris de xarxa i recursos d'Internet. La Figura 1 il·lustra l'arquitectura d'aquest sistema heretat, on una única ETL alimenta la base intermèdia (DB_AUX) tant des del gestor original (DB_PIO) com des de fonts externes, abans d'alimentar el Geoportal.

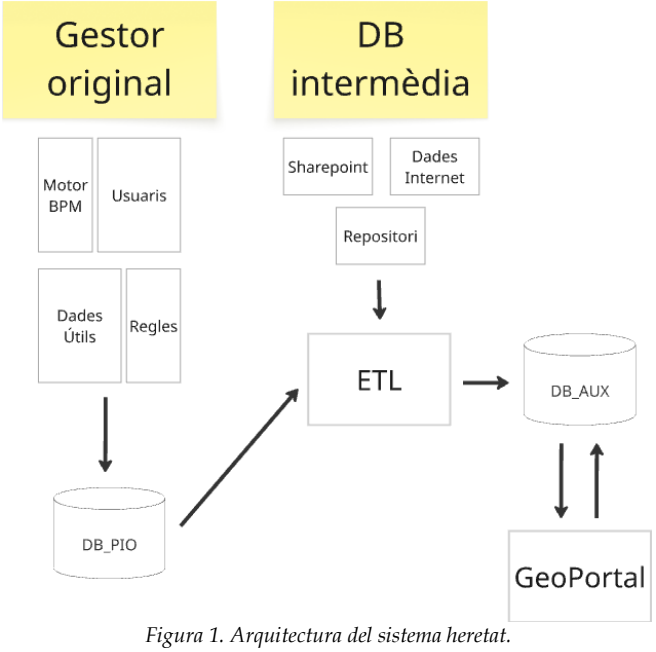


Figura 1. Arquitectura del sistema heretat.

Davant d'aquestes limitacions, l'AMB va impulsar un nou sistema de gestió amb una arquitectura dual (Figura 2). La utilització de dos motors de bases de dades diferenciats no respon a una decisió arbitrària, sinó a la naturalesa de cada base. La base principal (*DB_GESTOR*) replica l'estructura del nou gestor corporatiu desenvolupat per l'empresa externa i, per tant, n'hereta el motor MySQL. En

canvi, la base complementària (*DB_COMPLEMENTARI*) constitueix un disseny propi del projecte, per a la qual s'ha escollit PostgreSQL per la seva llibertat d'administració i pel seu suport natiu per a dades espacials mitjançant PostGIS, requisit imprescindible per a la gestió geogràfica del Geoportal. Així, tot i que ambdós motors comparteixen una sintaxi SQL àmpliament compatible, l'elecció de cadascun respon a un motiu concret: MySQL ve determinat pel sistema corporatiu, mentre que PostgreSQL aporta les capacitats geoespacials que el projecte requereix.

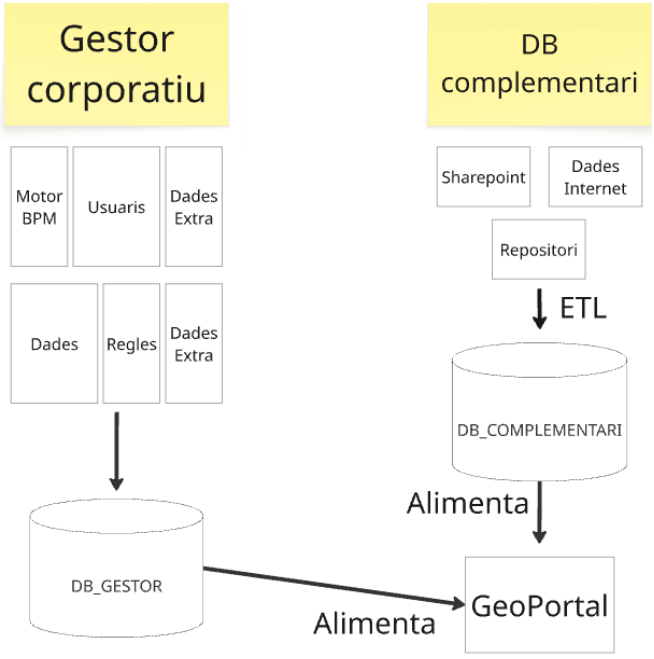


Figura 2. Arquitectura dual del sistema de dades.

2.2 Governança i Arquitectura de Dades

A partir de l'anàlisi del sistema heretat, s'han definit les regles de correspondència entre el sistema antic i el nou. Aquesta tasca ha anat més enllà d'un simple enllaç camp a camp, ja que ha requerit una reestructuració profunda de les dades per adaptar-les a la normalització que exigeix el nou gestor. Com a exemple pràctic, s'ha establert el criteri per unificar taules històriques fragmentades —com les de Finançament i Actuacions— dins les noves entitats mestres, determinant per a cadascuna l'exclusió de camps obsolets, els canvis de tipologia de dades i la reassignació de les claus foranes. El resultat és una matriu de correspondència que categoritza els camps en quatre tipus: camps d'integració nativa (valors categòrics ja definits al nou sistema que no requereixen migració), transferència directa (camps identificats i validats per a migració automatitzada), pendents d'interpretació funcional (camps detectats a l'origen el propòsit dels quals ha requerit clarificació amb l'empresa externa) i dades de nova entrada manual (buits d'informació on els requeriments actuals exigeixen dades que no es recopilaven prèviament).

Sobre la informació d'origen s'han definit tres capes de tractament: un criteri de filtratge i exclusió, un conjunt de regles de mapeig i homologació d'entitats, i un procés de *data cleansing*.

La primera capa és el criteri de filtratge i exclusió: només es migraran les actuacions que tinguin una planificació associada, s'exclouran les taules que el nou gestor ja carrega per defecte i es preservarà l'històric d'auditories exclusivament a la base de dades complementària per incompatibilitat de formats. La segona capa és el conjunt de regles de mapeig i homologació d'entitats, entre les quals destaca la unificació de les antigues classificacions d'assistències tècniques i obres sota una única entitat de contractes. La tercera capa és el *data cleansing* que comprèn la correcció de duplicitats toponímiques als noms dels municipis, la desagregació de registres múltiples on un sol camp contenia llistes de valors separats per comes i controls de qualitat que verifiquen la quadratura entre la suma de les anualitats i l'import total de cada projecte.

Paral·lelament, s'ha finalitzat el disseny de l'esquema relacional de la base de dades complementària (PostgreSQL) (vegeu Annex A1, Figura A.1). El model s'organitza en quatre mòduls funcionals. El **Mòdul de Control i Auditoria** inclou la taula *etl_controls*, que gestiona l'estat de les càrregues mitjançant *hashes* MD5 per evitar processaments redundants, i la taula *audits*, que centralitza el registre de canvis (valors anteriors i nous) per mantenir un històric auditable. El **Mòdul de Licitació i Contractació** articula el seguiment d'expedients a través de la taula *procurement_plans* (objecte, tipus de contracte, pressupost estimat), vinculada a la taula *lots* per gestionar les divisions contractuals, i la taula *resolution_registers* per monitorar l'estat administratiu i les dates clau del circuit intern. El **Mòdul Econòmic i d'Execució** constitueix el nucli de seguiment financer, pivotant sobre la taula *funding_allocations*, des de la qual es desplega el control d'anualitats aprovades i d'adjudicació, certificacions d'obra, modificacions d'import i contribucions de diferents fonts de finançament. Finalment, el **Mòdul de Gestió Operativa** incorpora la taula *timesheets* per portar un registre centralitzat del temps i els recursos tècnics que el personal de l'AMB dedica a cada cas.

Tot i que la base de dades complementària resideix en un motor independent, el seu disseny s'ha concebut per mantenir-se acoblada de manera feble (*loose coupling*) amb el gestor corporatiu. La integració entre ambdós sistemes es realitza mitjançant identificadors UUID5 [11] que actuen com a nexe d'unió, delegant les operacions de creuament al *backend* del Geoportal. Concretament, les taules *funding_allocations* i *procurement_plans* utilitzen els camps *cas* i *expedient* com a claus de vinculació amb el gestor extern, mentre que la taula *timesheets* es relaciona amb el directori actiu d'usuaris del gestor corporatiu. Aquest enfocament garanteix que l'AMB disposi d'una base de dades complementària àgil i escalable per a les seves necessitats específiques sense comprometre l'estabilitat del sistema transaccional central gestionat per l'empresa externa.

La taula *timesheets* registra la dedicació horària del personal tècnic de l'AMB, subjecta al Reglament General de Protecció de Dades (RGPD) [12]. El disseny de la base de dades complementària protegeix aquesta informació seguint el principi de minimització: actualment, mentre el gestor corporatiu extern no estigui disponible, la vincula-

ció es fa amb l'identificador d'usuari del personal; un cop desplegat el nou gestor, aquest camp se substituirà per un identificador de referència (UUID) al directori d'usuaris central, de manera que la informació personal identificable romandrà exclusivament sota la custòdia del sistema transaccional gestionat per l'empresa externa.

3 MIGRACIÓ

La migració de dades constitueix una de les fases més crítiques del projecte, ja que representa el pont entre el sistema heretat —basat en una arquitectura Oracle altament fragmentada amb dades distribuïdes en taules històriques poc normalitzades i amb un grau elevat de redundància— i el nou sistema de gestió, dissenyat sota un model relacional modern, normalitzat i orientat a la integritat referencial. L'objectiu d'aquesta fase no és simplement traslladar dades d'un origen a un destí, sinó transformar, validar i reestructurar la informació per adaptar-la al nou model relacional, tot preservant la traçabilitat històrica —garantida per la generació determinista d'identificadors UUID5— i la qualitat del resultat final, validada mitjançant els controls que es detallen a la Secció 6.

El context d'execució presentava una particularitat singular que va condicionar tot l'enfocament metodològic: el sistema destí és una plataforma desenvolupada per una empresa externa que ha acumulat retards significatius en el lliurament del producte. Atès que no es disposa d'accés a la base de dades de producció definitiva, s'ha optat per una estratègia consistent a sol·licitar a l'empresa externa l'esquema SQL original (definicions DDL, *Data Definition Language*) que defineix el nou sistema i replicar-ne l'estructura en un servidor MySQL allotjat al núvol (Aiven), creant així un entorn de migració funcional i totalment representatiu del sistema final. Aquesta decisió desacobla en gran mesura el progrés del nostre treball de les contingències del proveïdor extern. Quan es disposi de la base de dades definitiva, i sempre que se'n mantingui l'estructura, el redireccionament dels scripts requerirà principalment un canvi de cadena de connexió, sense necessitat de modificar la lògica de transformació. L'empresa externa ha confirmat que l'esquema es troba estabilitzat i que no es preveuen canvis estructurals de fons, més enllà de possibles addicions puntuals de camps, fet que fa raonable aquesta assumpció. L'esquema rebut conté 89 taules que estructuren tot el domini funcional del nou sistema.

Una migració eficient comença per identificar quines taules cal migrar i quines no, per no malbaratar esforços en taules que mai contindran dades del sistema antic. Per aquest motiu, s'han analitzat les 89 taules del sistema destí i classificat en quatre categories funcionals (vegeu Annex A2, Taula A.1). Les 13 taules de domini migrables inclouen, entre d'altres, *actions*, *services*, *users*, *programs*, *fundings*, *projects*, *certifications* i *economic_items*; cadascuna compta amb un *script* de migració dedicat amb lògica de transformació pròpia. Les 17 taules de referència (com action_types, cities, priorities o roles) actuen com a diccionaris de valors normalitzats, ja emplenats per l'empresa externa; es copien directament des d'una base de dades de referència mitjançant un patró idempotent que en cada execució trunca i recarrega la taula destí, garantint que els canvis als catàlegs es propaguin automàticament. Les 34 taules sense correspondència agrupen relacions que no existien en el sistema antic, com *action_user* i *role_user*, que registren la participació individual d'usuaris en cada actuació. El sistema antic operava amb usuaris compartits per departament, cosa que impedia reconstruir aquesta granularitat. Finalment, les 25 taules d'infraestructura pertanyen al *framework* Laravel [13] o al motor de regles de negoci i mai han de ser objecte de migració.

El *pipeline* s'ha dissenyat sota una arquitectura de tres capes (extracció, transformació i càrrega, ETL) amb una organització modular, orquestrada per *scripts/main.py* [14], que executa els passos en un ordre (vegeu Figura 3) que respecta les dependències referencials de la base de dades relacional: primer *users* i *services*, després *actions*, posteriorment *programs* i *fundings*, i finalment *projects*, *certifications*, *economic_items* i *anual_budget*s. Cada taula migrable disposa d'un mòdul propi amb estructura interna consistent: funcions d'extracció (Oracle i CSV per a execució *offline*), funcions de transformació (normalització de cadenes, *mapping* d'identificadors, generació d'UUIDs, càlculs derivats) i una funció d'inserció a MySQL via SQLAlchemy. Aquest disseny dual Oracle/CSV és deliberat, ja que permet executar la migració en mode connectat directament contra Oracle per a càrregues finals, o en mode *offline* a partir d'exportacions CSV quan no hi ha connectivitat amb la xarxa interna de l'AMB. Tots els scripts són idempotents: utilitzen la combinació (*action_id*, *code*) o equivalents com a clau natural per detectar registres ja inserits, i salten les insercions duplicades. Addicionalment, tots els registres incorporen un identificador UUID generat amb *uuid.uuid5()* a partir d'un *namespace* determinista, seguint l'estàndard RFC 9562 [11] de l'IETF, fet que garanteix que el mateix registre lògic obtingui sempre el mateix identificador entre execucions i faciliti la traçabilitat i les operacions d'actualització selectiva.

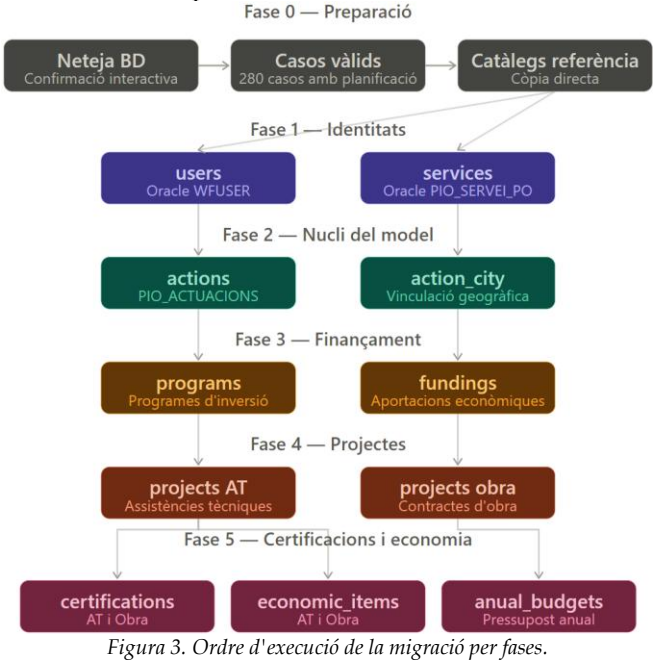


Figura 3. Ordre d'execució de la migració per fases.

Una de les decisions de disseny més rellevants ha estat migrar selectivament només una part de les dades, en lloc de traslladar-les totes. El sistema antic conté un volum considerable d'actuacions registrades al llarg dels anys, però una anàlisi qualitativa ha revelat que una part significativa presenta dades incompletes, incorrectes o senzillament obsoletes: actuacions sense planificació definida, registres antics amb camps crítics buits, o entrades que mai van arribar a tenir activitat real al sistema. Tal com assenyala Morris [15] en la seva teoria sobre la migració de sistemes heretats, traslladar dades sense un filtratge previ de qualitat només serveix per propagar el deute tècnic i comprometre la fiabilitat del nou sistema.

Davant d'aquesta realitat, es va decidir migrar únicament les actuacions amb planificació completa, un total de 280 casos, una reducció deliberada del 77,2% sobre les 1.227 actuacions del sistema original. Les actuacions no migrades corresponen a registres incomplets, obsolets o sense activitat real que no compleixen els criteris de qualitat establerts. La seva exclusió respon al requisit de no incorporar dades de baixa qualitat al nou sistema. Tot i així, aquesta informació no es perd: es conserva en una còpia de seguretat del sistema Oracle original, de manera que resta disponible per a una eventual consulta o recuperació futura. Aquesta decisió no respon a una limitació tècnica, sinó a un criteri de qualitat de dades que produeix dos beneficis directes: d'una banda, garanteix que el nou sistema parteix d'un *dataset* net, coherent i representatiu de l'activitat real, sense el deute tècnic del sistema antic; d'altra banda, exerceix una pressió organitzativa sobre els serveis gestors per regularitzar les actuacions noves sense planificació abans de la migració definitiva, millorant així la disciplina en l'entrada de dades.

Amb aquest enfocament, s'han migrat les 280 actuacions que compleixen el criteri de qualitat, amb les transformacions necessàries per adequar-les al nou model. Les taules no migrades queden agrupades en categories documentades i justificades: les taules de referència, que l'empresa externa proporciona ja emplenades amb els seus valors normalitzats; les taules sense correspondència, que s'ompliran progressivament amb l'ús de la nova aplicació; i les taules d'infraestructura del framework, que no contenen dades de negoci i, per tant, mai havien de ser objecte de migració. Aquesta delimitació de l'abast respon a un coneixement del domini i de les limitacions de tots dos sistemes. La tasca pendent de redireccionament, ja comentada, es resol amb un canvi de configuració.

4 Desenvolupament de Processos ETL

Un cop executada la migració, el treball se centra en la construcció de l'ecosistema d'automatització permanent, que comprèn els processos d'ETL i la capa de visualització i es gestiona íntegrament mitjançant el repositori de codi del projecte [16]. Aquesta secció se centra en l'ETL; la visualització es descriu a la Secció 5. És fonamental entendre la diferència conceptual entre l'ETL i la migració descrita a la secció anterior: mentre la migració és una operació puntual, l'ETL és un procés continu i automatitzat que s'executa periòdicament per mantenir sincronitzada la base de dades complementària amb les fonts d'origen, garantint que els indicadors, els informes i les visualitzacions del Geoportal reflecteixen sempre l'estat actualitzat de les actuacions de l'AMB. L'ETL actua exclusivament sobre la base de dades complementària; la base principal (DB_GESTOR) no forma part d'aquest flux, ja que s'alimenta directament del nou gestor corporatiu.

Aquest apartat descriu primer l'arquitectura del *pipeline* ETL (Secció 4.1) i, a continuació, les capes de robustesa, qualitat i portabilitat que el converteixen en un sistema de producció autogestionat (Secció 4.2).

4.1 Arquitectura del pipeline

El *pipeline* segueix el patró arquitectònic clàssic *Extract-Transform-Load*, prenent com a base les pràctiques d'integració de dades establertes per Kimball i Ross [17] i les tècniques específiques d'ETL descrites per Kimball i Caserta [18]. Cadascuna de les tres fases s'ha implementat com un mòdul independent dins el projecte, orquestrades per un controlador central que gestiona el flux d'execució, la gestió d'errors i la coordinació entre fases mitjançant un mòdul *PipelineConfig* que centralitza tota la configuració. Per garantir la seguretat de les execucions, el *pipeline* disposa de dos entorns diferenciats —preproducció i producció— cadascun connectat a la seva pròpia base de dades, de manera que qualsevol canvi pot ser validat en preproducció abans de propagar-se al sistema productiu.

La fase d'**Extracció** s'encarrega d'obtenir dades de tres fonts radicalment diferents, cadascuna amb la seva pròpia complexitat tècnica. La primera és la base de dades Oracle del sistema antic, accessible via la llibreria *oracledb*, des d'on s'extreuen les dades històriques d'actuacions, finançaments i certificacions. La segona són els fitxers Excel emmagatzemats a SharePoint, que requereixen autenticació corporativa via *Office365-REST-Python-Client* i contenen la planificació actualitzada manualment pels gestors. La tercera són els serveis ArcGIS Portal i ArcGIS Online, accessibles via arcgis Python API, que aporten la geometria territorial de cada actuació.

La fase de **Transformació** és on s'aplica el coneixement del domini funcional. Els mòduls d'aquesta fase realitzen normalitzacions de cadenes (eliminació d'espais, harmonització d'accents, *mapping* de valors equivalents), càlculs derivats (totals agregats, anualitats, percentatges d'execució), generació d'identificadors deterministes UUID5 calculats a partir de combinacions de camps clau, i la implementació de la lògica de negoci específica del projecte. El submòdul *financament_programs/* conté els transformadors específics per a cadascuna de les entitats del model de finançament (*allocations*, *contributions*, *annuities*, *certifications*, *modifications*, *mapping_allocations*). Com que les dades d'origen presenten una estructura jeràrquica en què només la fila pare conté l'identificador *cas* i les files filles el deixen buit, s'implementa una operació de *forward-fill*: el valor de *cas* es propaga cap avall a totes les files filles fins a trobar un nou pare, de manera que cada fila queda correctament associada al seu expedient.

La fase de **Càrrega** diposita els *dataframes* (taules de dades en memòria de Pandas) transformats a la base de

dades destí (PostgreSQL) amb els controls necessaris per garantir la integritat referencial. Els scripts utilitzen *psycopg2*, l'adaptador de PostgreSQL per a Python, i implementen estratègies d'inserció idempotents: segons si el registre ja existeix a la base de dades —comprovat per la seva clau natural—, decideixen entre actualitzar-lo (UPDATE) o inserir-lo de nou (INSERT), sense generar mai duplicats. La càrrega s'organitza respectant l'ordre de dependències del model: primer les taules pare (*funding_allocations*), després les filles directes (*mapping_allocations*, *funding_contributions*), i finalment les entitats que depenen de les filles (*annuities*, *certifications*, *modifications_adjudication*).

A més del flux principal d'actuacions, el *pipeline* integra altres dominis funcionals provinents de SharePoint, com el Pla de Contractacions —materialitzat a les taules *procurement_plans*, *lots* i *resolution_registers*— i els Acords de Govern. Aquests fitxers s'incorporen al cicle horari d'actualització i se sotmeten al mateix sistema de detecció de canvis per *hash* i d'auditoria (descrits a la Secció 4.2). Cal assenyalar que el seguiment d'acords (*tracking_agreements*) també s'ha desenvolupat completament, si bé s'ha detectat una inconsistència a les dades d'origen que impedeix generar files úniques de forma fiable. Per aquest motiu, i un cop comunicada la incidència a l'equip responsable, la integració al cicle automàtic resta desactivada fins a la resolució del problema.

4.2 Robustesa, qualitat i portabilitat

Sobre aquesta base arquitectònica, s'hi han desplegat diverses capes que converteixen el *pipeline* en un sistema de producció autogestionat. La primera capa és el sistema de detecció de canvis basat en *hashes* MD5. El sistema calcula el *hash* MD5 [19] de cada *dataframe* extret i el compara amb el *hash* emmagatzemat a la taula *etl_controls*. Si els *hashes* coincideixen, la transformació i la càrrega de la taula s'omet completament; si difereix, el *pipeline* executa el processament i actualitza el *hash* registrat. Aquesta optimització redueix els temps d'execució i ofereix traçabilitat de quina taula es va modificar i quan.

Complementant aquest control, el *pipeline* incorpora un **sistema d'auditoria** per columnes que registra cada modificació individual a la taula *audits*: el nom de la taula afectada, l'identificador del registre modificat, la columna alterada, el valor antic, el valor nou i el *timestamp* del canvi. Aquest sistema és fonamental en un context on diferents tècnics modifiquen manualment dades als fitxers Excel d'origen: davant de qualsevol incidència sobre la integritat d'una dada concreta, és possible reconstruir tot l'historial de modificacions d'aquella dada.

Per a la **validació de qualitat de dades** s'ha integrat un mòdul dedicat que executa una bateria de comprovacions abans i durant la càrrega. Les regles s'han dissenyat conjuntament amb els tècnics gestors per detectar les inconsistències: coherència geogràfica (que el municipi del finançament coincideixi amb el registrat al sistema), quadratura pressupostària (amb una tolerància configurable d'1€ per absorbir arrodoniments) i saturació de certificacions (assegurant que la suma de certificacions emeses per a una actuació no superi mai el total disponible, evitant sobrecàrregues pressupostàries). Quan una validació falla, el *pipeline* no atura l'execució —això podria deixar el sistema en un estat inconsistent— sinó que registra la incidència en un *log* estructurat i envia una notificació automàtica al responsable, de manera que el sistema es manté operatiu mentre s'aborden les anomalies.

El sistema de **notificacions** distingeix quatre tipus de missatges: *alertes crítiques* davant errors d'execució, *recordatoris de termini* (amb 30 i 14 dies d'antelació), *alertes de mapping* quan apareixen nous registres pendents de revisió manual i *resums setmanals personalitzats* per a cada servei d'obres. Els correus es generen amb una plantilla HTML corporativa i s'envien via el servidor SMTP intern de l'AMB. Finalment, s'ha implementat el **patró Adaptador** per a les connexions a bases de dades. La seva finalitat és aïllar tota la lògica de transformació i càrrega del *pipeline* de l'origen i el destí concrets de les dades: al seu nivell, els transformadors no saben (ni necessiten saber) si les dades venen d'Oracle, MySQL, PostgreSQL o un fitxer CSV. Aquesta arquitectura d'acoblament feble és el que fa possible el canvi de la base de dades de gestor un cop l'empresa externa lliuri el seu producte definitiu: l'únic canvi necessari serà actualitzar la configuració de connexió al mòdul *config*, mentre que tota la lògica de transformació, validació, generació d'UUIDs, càlcul de *hashes* i auditoria continuarà funcionant sense cap modificació.

5 VISUALITZACIÓ I MODELATGE PREDICTIU

Un cop l'ETL ha garantit la qualitat i l'estructuració de la informació, la capa de presentació constitueix l'etapa final de l'explotació de dades. L'objectiu d'aquesta capa és transformar les dades processades en informació estratègica interpretable pels tècnics i gestors de l'AMB a través del Geoportal. El present projecte se centra en el desenvolupament de la capa de dades i la llibreria d'informes (helpers_lib); el *backend* i el *frontend* del Geoportal corresponen a l'equip de sistemes corporatius. La feina desenvolupada en aquest projecte garanteix que la interfície disposi d'una API de dades estructurada i de models predictius llestos per ser renderitzats en temps real.

Aquesta capa s'estructura en dos blocs: l'automatització de la generació d'informes institucionals (Secció 5.1) i un model predictiu que estima el risc de retard en el compliment de les subfases de planificació (Secció 5.2).

5.1 Automatització del reporting institucional

S'ha desenvolupat un sistema de generació automàtica d'informes institucionals. Històricament, la producció d'aquests informes requeria que els tècnics de l'AMB realitzessin extraccions manuals de dades disperses en múltiples fulls de càlcul Excel, un procés repetitiu, propens a errors i que consumia moltes hores de treball mensual. Amb el nou flux queda automatitzat: el *backend* Python consulta directament les entitats del mòdul econòmic de la base de dades complementària, construeix els *dataframes* i els passa a mòduls de renderització independents per a cada format de sortida (Figura 4). Quan el gestor corporatiu entri en producció, el *backend* també el podrà consultar directament com a font de dades.

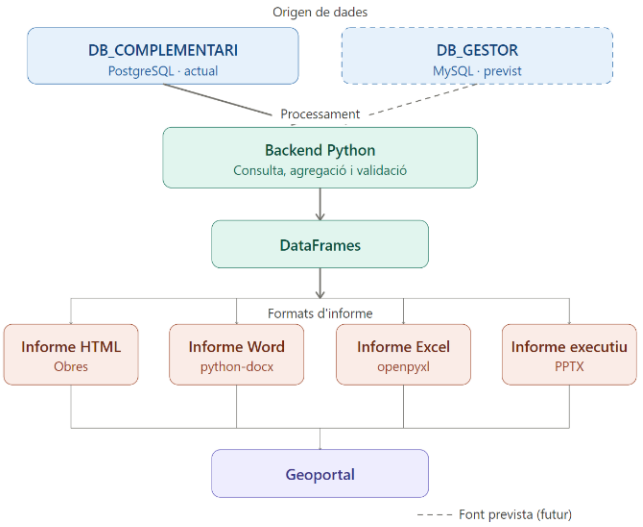


Figura 4. Arquitectura de la capa de presentació.

La capa de presentació segueix un patró de separació de responsabilitats estricte: la lògica de consulta, agregació i validació de dades s'executa íntegrament al *backend*, desacoblada dels mòduls de renderització. Per evitar la duplicació de codi entre l'ETL i el *backend* del Geoportal, les funcions compartides s'han centralitzat en una llibreria interna instal·lable (*helpers_lib* [20]): així, qualsevol actualització es propaga automàticament als dos sistemes i es manté una única font de codi. Tots els informes s'integren al Geoportal amb un motor de filtratge que permet segmentar la informació per diversos criteris.

El sistema exposa quatre formats de sortida, cadascun adaptat a un perfil d'usuari i cas d'ús diferent:

1) **Informe HTML d'obres**: Orientat al seguiment visual d'actuacions individuals (vegeu Annex A3, Figura A.2). Incorpora una seqüència de *fallback* per obtenir el recurs visual més adequat: realitza primer una petició REST al banc d'imatges corporatiu de l'AMB (BIAM); si no existeix registre fotogràfic, sol·licita la capa topogràfica de l'àmbit geogràfic corresponent; si tampoc està disponible, el sistema genera pictogrames representatius de la tipologia d'actuació.
2) **Informe Word**: Generat amb *python-docx* [21], cobreix totes les tipologies d'actuació amb la informació econòmica i administrativa completa. Exportable a PDF seguint l'estil corporatiu.
3) **Informe Excel**: Generat amb *openpyxl* [22], és l'equivalent tabular del Word i incorpora les dates clau de cada *funding_allocation*.
4) **Informe executiu**: Dissenyat per als òrgans de govern i ajuntaments (vegeu Annex A3, Figura A.3 i Figura A.4). A nivell tècnic, utilitza una plantilla corporativa PPTX on substitueix programàticament les variables i els gràfics estàtics per valors reals en temps d'execució. A continuació, aquest document es renderitza i s'exposa en format HTML, permetent la visualització nativa i integració directa dins del *frontend* de la plataforma, sense necessitat de descàrregues externes.

5.2 Model Predictiu de KPIs

Com a component avançat de la capa d'explotació, s'ha dissenyat un model d'aprenentatge automàtic per predir la probabilitat que cada subfase de planificació d'una actuació es completi dins del termini estimat. L'objectiu és proporcionar als gestors un indicador clau de rendiment (KPI, *Key Performance Indicator*) quantitatiu del risc de retard, calculat per a cada subfase activa, de manera que la presa de decisions en la gestió del portafolis pugui anticipar-se als problemes en lloc de reaccionar-hi a posteriori. El model no pretén substituir el criteri tècnic dels professionals, sinó actuar com una eina de suport a la decisió (*Decision Support System*) que assenyali on posar el focus d'atenció. Aquest enfocament es materialitza al *frontend* del sistema (Figura 5), on les probabilitats calculades es tradueixen en un codi visual intuïtiu de semàfors que facilita el seguiment actiu de l'estat de les actuacions.

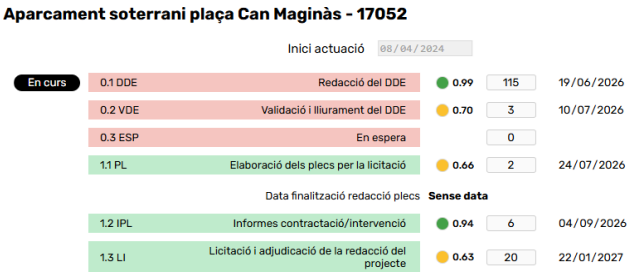


Figura 5. Indicadors de risc de retard integrats a la interfície del Geoportal per a l'actuació Aparcament soterrani plaça Can Maginàs (17052).

Durant l'auditoria del model es va detectar un problema metodològic greu que cal documentar amb transparència: el *dataset* original presentava un *data leakage* estructural que invalidava les mètriques aparents del model. L'anàlisi va revelar tres problemes encadenats. En primer lloc, una fuga d'informació en l'estimació: el registre de modificacions contenia 11.072 files, però en els casos sense entrada al registre, el camp d'estimació es sobreescrivia automàticament amb la duració real observada, fet que feia que el model rebés com a *feature* el mateix valor que intentava predir. En segon lloc, dades introduïdes a posteriori: una anàlisi temporal va revelar que les dades del registre de modificacions s'havien introduït amb una mediana de 247 dies després de la finalització de la subfase corresponent, de manera que el que es presentava com a estimació prèvia era en realitat un registre històric capturat a posteriori. En tercer lloc, un *split* de validació incorrecte: la partició *train/validation* es feia aleatòriament a nivell de fila, ignorant que múltiples files corresponien a la mateixa actuació, fet que provocava solapament massiu entre conjunts. L'efecte combinat era una AUC (*Area Under the Curve*) aparent de 0,974 —i en les primeres iteracions propera a 1,0 amb un Brier de ~0,002— que no reflectia cap capacitat predictiva real sinó un artefacte del disseny experimental. Aquest fenomen, documentat per Kaufman et al. [23], es coneix com a *target leakage* i és precisament el tipus de biaix que només una auditoria sistemàtica del *pipeline* complet permet identificar.

A partir d'aquesta detecció, s'ha reconstruït la metodologia d'entrenament aplicant les pràctiques estàndard de

l'aprenentatge automàtic supervisat.

El primer canvi ha estat un filtratge estricte del *dataset*: només s'inclouen les files que disposen d'una modificació real registrada amb anterioritat a la finalització de la subfase, reduint el conjunt utilitzable a només 457 casos reals. El segon ha estat la implementació d'una validació creuada amb agrupació estricta a nivell d'actuació (GroupKFold [24] per obra): totes les files corresponents a una mateixa obra cauen exclusivament en el conjunt d'entrenament (*train*) o en el de validació (*validation*), mai en ambdós simultàniament.

Pel que fa a l'arquitectura, s'ha entrenat un model de *gradient boosting* (una tècnica que combina múltiples arbres de decisió) mitjançant la llibreria LightGBM [25]. L'elecció d'aquest model respon a les característiques del problema: un conjunt de dades reduït (457 casos) amb un nombre elevat de variables. En aquestes condicions, els models de *gradient boosting* basats en histogrames com LightGBM solen oferir un millor rendiment i una menor tendència al sobreajust que altres alternatives avaluades, com Random Forest o XGBoost. El model incorpora restriccions monotòniques sobre les variables d'estimació —garantint que a major estimació de temps, major probabilitat de compliment— i *early stopping* de 50 rondes (aturada de l'entrenament si el rendiment no millora durant 50 iteracions consecutives, per evitar el sobreajust). Finalment, sobre la probabilitat bruta del model s'aplica una capa de penalització basada en els rangs històrics reals per subfase: si l'estimació queda per sota del mínim històric observat, s'aplica una penalització quadràtica agressiva; si l'estimació és generosa, s'aplica un bonus suau.

Un cop corregit el *data leakage* i aplicada aquesta metodologia d'entrenament, les mètriques obtingudes reflecteixen el rendiment real del model sobre actuacions completament noves, tal com es detalla a la Taula 1. Aquestes mètriques s'han calculat sobre el conjunt de validació mitjançant l'esquema GroupKFold descrit anteriorment, de manera que mesuren el rendiment del model sobre obres que no ha vist durant l'entrenament.

Mètrica d'avaluació	Resultat
Accuracy	0,735
AUC-ROC	0,756 ± 0,048
F1-Score	0,641
Baseline (classe majoritària)	0,613

Taula 1. Mètriques de rendiment del model predictiu LightGBM sobre el conjunt de validació.

Com s'observa a les mètriques, d'una banda, l'AUC de 0,756 reflecteix una capacitat discriminativa real (recordem que AUC = 0,5 equival a un classificador aleatori i AUC = 1 a un classificador perfecte). D'altra banda, l'accuracy supera el *baseline* de la classe majoritària en 12,2 punts percentuals.

Cal contextualitzar honestament aquestes limitacions. Els 457 casos amb modificacions reals registrades són insuficients per entrenar un model amb la robustesa que assolirien sistemes equivalents en àmbits més madurs. Aquesta limitació no és atribuïble al disseny del model, sinó al fet que el registre estructurat de modificacions d'estimacions és una pràctica recent dins l'organització, introduïda com a part del present projecte. Cada nou cicle d'enriquiment incorpora dades addicionals al *pool* d'entrenament potencial, fet que farà evolucionar el model de manera natural mitjançant reentrenaments periòdics, sense necessitat de redisseny estructural.

El valor d'aquest mòdul dins del projecte no rau únicament en les mètriques, sinó en haver construït una infraestructura predictiva metodològicament sòlida i rigorosament validada, sobre la qual el rendiment millorarà de manera orgànica a mesura que les dades de l'organització s'acumulin. En un context on la pràctica habitual hagués estat acceptar les mètriques fictícies inicials com a vàlides, la decisió d'auditar, identificar i corregir els biaixos metodològics constitueix una aportació tècnica diferencial del treball.

6 TEST I VALIDACIÓ

La fase de test i validació del projecte no s'ha concebut com una etapa final aïllada, sinó com un procés continu i integrat dins de cada cicle de desenvolupament tecnològic. En sistemes que processen dades crítiques per a la gestió pública, qualsevol error no detectat es propaga al llarg de tot el *pipeline* i pot acabar afectant decisions de gestió, informes institucionals o assignacions pressupostàries. Atesa la complexitat arquitectònica del sistema desenvolupat, s'ha dissenyat una estratègia de validació multinivell estructurada en cinc nivells complementaris.

El primer nivell és la **validació de la migració i integritat referencial**. La validació quantitativa ha confirmat, comparant el recompte de registres entre origen i destí, que les 280 actuacions seleccionades s'han migrat íntegrament i sense pèrdues. Paral·lelament, s'ha realitzat una anàlisi dels casos descartats classificant cada cas segons el criteri d'exclusió que li corresponia i comprovant que la totalitat dels descartats quedava coberta per algun criteri legítim (absència de planificació, dades obsoletes, registres incomplets), sense que cap registre quedés exclòs per un error de l'script de migració. A nivell d'integritat referencial, cada clau forana action_id referenciada per les taules dependents (*fundings*, *projects*, *certifications*, *economic_items* i *action_city*) ha estat verificada contra la taula *actions* per garantir la consistència de totes les relacions, i s'ha implementat una validació de consistència semàntica sobre camps clau (imports econòmics dins de rangs raonables, dates coherents amb el cicle de vida d'una actuació, identificadors únics correctament generats mitjançant UUID5 determinista).

El segon nivell són les **proves unitàries del codi ETL**. S'ha desenvolupat una bateria de 35 tests organitzats en quatre mòduls: tests del *pipeline* ETL i el seu controlador (16), tests de validació de qualitat de dades de finançament (8), tests de generació d'informes institucionals (6) i tests de validació estadística del model predictiu (5, incloent-hi 1 marcat com *xfail* —test que es preveu que falli fins a aplicar una correcció pendent al *dataset_init*). Tots els mòduls utilitzen *pytest* en combinació amb *unittest.mock* per simular dependències externes i *dataframes* sintètics de Pandas per controlar les dades d'entrada. Els

tests cobreixen: el controlador del *pipeline* (instanciació, detecció de canvis per *hash*, robustesa davant atributs faltants), la resistència a errors (*timeouts*, connexions perdudes, autenticació fallida), els fluxos condicionals i casos límit (*dataframes* buits, mida mínima), els tres *checks* de qualitat de finançament (coherència de municipi, quadratura pressupostària i saturació de certificacions), els formats d'informe (Word, Excel, PPTX i PDF), i la validació estadística del model (GroupKFold sense solapament, restricció monotònica i penalització per estimació fora de rang). Addicionalment, s'ha implementat un test d'integració end-to-end que executa el *pipeline* complet amb totes les dependències externes simulades. La bateria completa de 35 tests es supera correctament en l'entorn del projecte amb totes les dependències instal·lades.

El tercer nivell és la **validació de qualitat de dades** (*Data Quality*) executada en temps real pel *pipeline*. Aquest tipus de validació no es pot resoldre amb tests estàtics: requereix comprovacions dinàmiques sobre les dades reals a cada execució. Les regles, descrites en detall a la Secció 4, actuen com a darrera línia de defensa abans que les dades arribin a la base de dades complementària. Quan una validació falla, el *pipeline* no atura l'execució sinó que registra la incidència de manera estructurada i genera notificacions automàtiques als responsables corresponents, una estratègia de degradació controlada que és la pràctica recomanada en sistemes ETL de producció.

El quart nivell és la **validació estadística del model predictiu**. La metodologia s'ha explicat a la Secció 5. Ha estat precisament aquest nivell el que ha permès detectar i corregir un *data leakage* estructural que invalidava les mètriques aparents del model original. La metodologia final incorpora les millors pràctiques de validació: GroupKFold per actuació (garantint que el model mai vegi en validació cap registre d'una actuació que hagi vist en entrenament), càlcul d'estadístics exclusivament sobre el *train* (evitant fuites d'informació entre conjunts), optimització del *threshold* sobre validació, i *reporting* de mitjanes amb desviació estàndard entre *folds* (oferint una visió honesta de la variabilitat estadística en lloc d'amagar-la sota un únic número agregat).

El cinquè i últim nivell és la **validació funcional i d'interfície** (*User Acceptance Testing* [26]), que connecta l'enginyeria del projecte amb el valor de negoci percebut pels gestors de l'AMB. Els quatre formats d'informe s'han generat amb diferents combinacions de filtres, i els gestors de l'AMB n'han revisat les sortides, confirmant tant la correcció de les dades com l'adequació de les visualitzacions a la identitat corporativa de l'AMB. La cadena de prioritats visuals de l'informe HTML d'obres (fotografia BIAM → topografia → pictograma) s'ha provat en els tres escenaris possibles: actuacions amb fotografia disponible, actuacions sense fotografia però amb capa topogràfica, i actuacions sense cap dels dos, que recauen en el pictograma. El motor de filtratge multidimensional s'ha validat amb diverses combinacions de paràmetres, comprovant que els resultats coincideixen amb les expectatives dels gestors. Finalment, s'ha verificat que les prediccions del model i els informes generats es consumeixen correctament des del *frontend* del Geoportal, on ja estan integrats i operatius. Pel que fa al rendiment, la generació dels informes presenta temps de resposta adequats: el cas més costós (el municipi de Badalona, amb un nombre elevat d'actuacions sense fotografia que requereixen renderitzar la capa topogràfica) triga uns 20 segons, mentre que la resta de casos es resolen en temps notablement inferiors. L'optimització d'aquest procés de renderitzat queda identificada com a possible línia de millora futura.

La combinació d'aquests cinc nivells conforma una estratègia de validació multinivell que cobreix des del codi fins a l'usuari final. Més enllà del nombre de proves executades, és aquesta cobertura a diferents nivells la que aporta solidesa al conjunt del sistema.

7 Conclusions

Partint d'un ecosistema de dades fragmentat i amb problemes estructurals d'integritat, aquest treball ha culminat en una infraestructura tècnica robusta que cobreix els tres eixos d'enginyeria plantejats inicialment: la migració del sistema heretat, el desenvolupament d'un *pipeline* ETL d'integració contínua i la capa d'explotació de dades amb informes automatitzats i model predictiu.

La migració ha consolidat les 280 actuacions amb planificació completa. La resta, que constituïa el 77,2% del sistema original, corresponia a registres incomplets o obsolets que es van excloure deliberadament per criteri de qualitat, no per limitació tècnica. El *pipeline* ETL es troba operatiu en producció i alimenta diàriament la base de dades complementària, integrant dades de tres fonts heterogènies (Oracle, SharePoint i ArcGIS), executant validacions automàtiques de qualitat, mantenint un sistema d'auditoria per columnes i emetent notificacions proactives als responsables. La capa d'explotació incorpora quatre formats d'informe institucional automatitzats que substitueixen un procés manual que prèviament consumia desenes d'hores mensuals, i un model predictiu de compliment de subfases amb una AUC de 0,756, metodològicament sòlid després d'haver detectat i corregit un *data leakage* estructural al *pipeline* d'entrenament original.

Més enllà dels lliurables tècnics, el treball aporta un valor institucional rellevant. Per primera vegada, el departament d'Espai Públic disposa d'una font única de veritat per a la informació econòmica i de seguiment d'obres, eliminant les duplicitats i les inconsistències que durant anys havien dificultat la presa de decisions. Els informes que abans es generaven manualment ara es produeixen sota demanda amb dades permanentment actualitzades, i el model predictiu —tot i les seves limitacions actuals— constitueix el primer sistema d'aprenentatge automàtic desplegat sobre les dades de planificació de l'AMB, obrint la porta a una cultura de gestió basada en l'evidència.

Des d'un punt de vista metodològic, aquest projecte ha posat de manifest aprenentatges clau. El cas més destacat ha estat la detecció i correcció del *data leakage* al model predictiu: el sistema original presentava mètriques aparentment excel·lents (AUC 0,974) que no reflectien cap capacitat predictiva real. Identificar aquest biaix i reconstruir la metodologia d'entrenament ha estat un dels apre-

nentatges principals del treball. Paral·lelament, l'ús del patró Adaptador ha permès avançar amb autonomia malgrat els retards del proveïdor extern.

En síntesi, els objectius plantejats inicialment s'han assolit. S'ha realitzat l'auditoria i definit el model de governança de dades; s'ha dissenyat la base de dades complementària; s'han desenvolupat els procediments de migració des del sistema heretat; s'han implementat els processos ETL d'actualització periòdica; s'ha alimentat el Geoportal amb les noves dades, els gràfics automatitzats i el model de risc; s'ha desenvolupat el model predictiu de compliment de subfases; i s'ha validat la integritat i qualitat de les dades resultants.

Com a feina a fer a futur, es preveu fer un seguiment de l'evolució de les mètriques del model (*accuracy* i AUC-ROC) a mesura que el conjunt d'entrenament creixi més enllà dels 457 casos actuals, per validar si l'arquitectura de reentrenament periòdic es tradueix en una millora real del rendiment predictiu. La modularitat del *pipeline* ETL facilita l'extensió a nous dominis funcionals — com la integració amb nous serveis territorials o futurs sistemes corporatius— sense haver de tocar els fonaments del sistema. Finalment, la base de dades complementària, concebuda com una extensió flexible del gestor corporatiu, queda preparada per absorbir noves entitats —com ara nous indicadors i mètriques de seguiment, o noves estructures de dades derivades dels canvis normatius habituals en l'administració pública— a mesura que les necessitats operatives de l'AMB ho requereixin.

AGRAÏMENTS

Vull expressar el meu sincer agraïment al tutor d'aquest Treball de Final de Grau, Enric Martí Gòdia, per la seva orientació i disponibilitat al llarg de tot el procés.

Vull agrair també a l'Àrea Metropolitana de Barcelona (AMB) l'oportunitat de desenvolupar aquest projecte en un context real i d'alt impacte institucional. Una menció especial a Jose Carlos Sánchez, que ha exercit com a supervisor del dia a dia durant tots aquests mesos i la implicació i el coneixement tècnic del qual han estat determinants. Gràcies també a Javier Duarte, tutor acadèmic per part de l'AMB, per la confiança dipositada des del primer moment.

Finalment, a la família i amics gràcies pel suport incondicional durant aquests mesos de dedicació intensa, i als companys del Grau en Enginyeria de Dades les hores compartides i el coneixement transmès al llarg dels darrers anys.

BIBLIOGRAFIA

[1] https://docs.oracle.com/en/database/, documentació oficial del motor de base de dades Oracle Database (data d'últim accés: febrer 2026).

[2] https://learn.microsoft.com/en-us/sharepoint/, documentació oficial de la plataforma col·laborativa SharePoint de Microsoft per a la gestió documental (data d'últim accés: juny 2026).

[3] https://doc.arcgis.com/en/, documentació oficial de la plataforma SIG ArcGIS d'Esri per a la gestió de dades geoespacials (data d'últim accés: juny 2026).

[4] Àrea Metropolitana de Barcelona (AMB). Geoportal. Aplicatiu web intern corporatiu de gestió espacial (Accés restringit).

[5] https://dev.mysql.com/doc/, documentació oficial del sistema de gestió de bases de dades relacionals de codi obert de MySQL (data d'últim accés: febrer 2026).

[6] https://www.postgresql.org/docs/, documentació oficial del motor de base de dades PostgreSQL (data d'últim accés: maig 2026).

[7] https://pandas.pydata.org/docs/, documentació oficial de la llibreria Pandas per a l'anàlisi i manipulació de dades en Python (data d'últim accés: abril 2026).

[8] https://github.com/vgrem/Office365-REST-Python-Client, documentació oficial de la llibreria Office365-REST-Python-Client per a l'accés a SharePoint i Microsoft Graph des de Python (data d'últim accés: juny 2026).

[9] https://developers.arcgis.com/python/, documentació oficial de l'ArcGIS API for Python d'Esri per a l'accés i la gestió de dades geoespacials (data d'últim accés: juny 2026).

[10] https://git-scm.com/doc, documentació oficial del sistema de control de versions Git (data d'últim accés: juny 2026).

[11] https://datatracker.ietf.org/doc/html/rfc9562, documentació oficial de l'estàndard IETF per a la generació d'identificadors únics universals UUID (data d'últim accés: maig 2026).

[12] https://eur-lex.europa.eu/eli/reg/2016/679/oj, Reglament (UE) 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques pel que fa al tractament de dades personals (RGPD) (data d'últim accés: maig 2026).

[13] https://laravel.com/docs, documentació oficial del framework de desenvolupament web Laravel basat en PHP (data d'últim accés: juny 2026).

[14] https://github.com/isanchezAMB/manager-migration, repositori de codi a GitHub amb els scripts de migració desenvolupats per al projecte (data d'últim accés: maig 2026).

[15] J. Morris, Practical Data Migration, 2a ed., BCS, The Chartered Institute for IT, 2012.

[16] Repositori privat de GitHub: Administració i control de versions del codi font (ETL, arquitectura de dades i mòduls de visualització). Àrea Metropolitana de Barcelona (AMB), juny 2026.

[17] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, John Wiley & Sons, 2013.

[18] R. Kimball, J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data, John Wiley & Sons, 2004.

[19] https://datatracker.ietf.org/doc/html/rfc1321, documentació oficial de l'estàndard IETF que defineix l'algorisme de resum MD5 (data d'últim accés: juny 2026).

[20] Repositori privat de GitHub: Llibreria interna de funcions compartides (helpers_lib) per als processos ETL i el backend. Àrea Metropolitana de Barcelona (AMB), juny 2026

[21] https://python-docx.readthedocs.io, documentació oficial de la llibreria python-docx per a la creació i edició de documents Word amb Python (data d'últim accés: juny 2026).

[22] https://openpyxl.readthedocs.io, documentació oficial de la llibreria openpyxl per a la lectura i escriptura de fitxers Excel amb Python (data d'últim accés: juny 2026).

[23] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 6, no. 4, pp. 1-21, 2012.

[24] https://scikit-learn.org/stable/, documentació oficial de la llibreria scikit-learn d'aprenentatge automàtic per a Python (data d'últim accés: juny 2026).

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3146-3154.

[26] I. Sommerville, Software Engineering, 10a ed., Pearson, 2015.

ANNEXOS

A1. Diagrama Relacional de la Base de Dades Complementària

A continuació es mostra l'esquema relacional de la base de dades complementària implementada en PostgreSQL. Per motius de llegibilitat, se'n presenta una versió simplificada que inclou únicament les entitats i els camps clau. S'hi poden observar els quatre mòduls funcionals descrits a la memòria, així com les relacions d'integritat referencial i la gestió d'identificadors UUID.

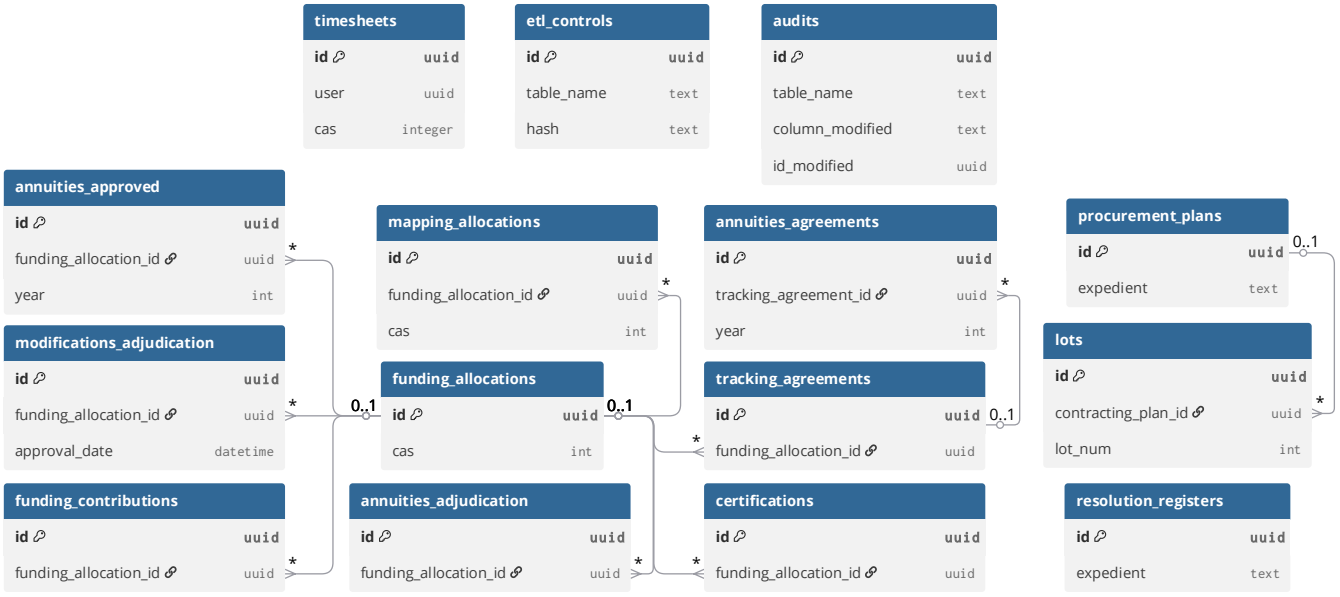


Figura A.1. Diagrama relacional simplificat de la base de dades complementària.

Tot seguit es descriu breument la funció de cada taula, agrupada pels quatre mòduls funcionals introduïts a la Secció 2.2:

Mòdul de Control i Auditoria:

- *audits*: registra cada modificació (camp, valor anterior, valor nou i *timestamp*) per a un historial complet de canvis.
- *etl_controls*: emmagatzema el hash de cada càrrega per ometre el reprocessament si no hi ha hagut canvis.

Mòdul de Licitació i Contractació:

- *procurement_plans*: plans de contractació d'expedients (objecte, tipus de contracte, pressupost).
- *lots*: divideix cada pla de contractació en lots, per a contractacions parcials.
- *resolution_registers*: estat administratiu i dates clau del circuit intern de l'expedient.

Mòdul Econòmic i d'Execució:

- *funding_allocations*: taula mare del mòdul; cada fila és un cas únic (camp cas) i punt de referència de la resta.
- *mapping_allocations*: registres de funding_allocations pendents de revisió manual quan apareix un valor sense correspondència coneguda.
- *annuities_approved*: anualitats pressupostàries aprovades per a cada finançament.
- *annuities_adjudication*: anualitats ja adjudicades per a cada finançament.
- *annuities_agreements*: desglossa per any l'import dels convenis de seguiment associats.
- *funding_contributions*: desglossa l'import d'un finançament entre les fonts que hi contribueixen.
- *certifications*: pagaments periòdics d'una obra (import i data de cada certificació).
- *modifications_adjudication*: modificacions d'import sobre una adjudicació existent, amb data d'aprovació.
- *tracking_agreements*: seguiment dels convenis de finançament amb altres administracions.

Mòdul de Gestió Operativa:

- *timesheets*: hores que el personal tècnic de l'AMB dedica a cada cas.

A2. CLASSIFICACIÓ DE LES TAULES DEL SISTEMA DESTÍ

A continuació es presenta la classificació completa de les 89 taules que conformen l'esquema del nou gestor corporatiu, agrupades en quatre categories funcionals segons la seva estratègia de migració. Aquesta anàlisi prèvia és el que permet delimitar amb precisió l'abast real de la migració, distingint les taules que requereixen un tractament dedicat de les que es carreguen automàticament, no tenen correspondència amb el sistema antic o pertanyen a la infraestructura del framework i, per tant, queden excloses del procés.

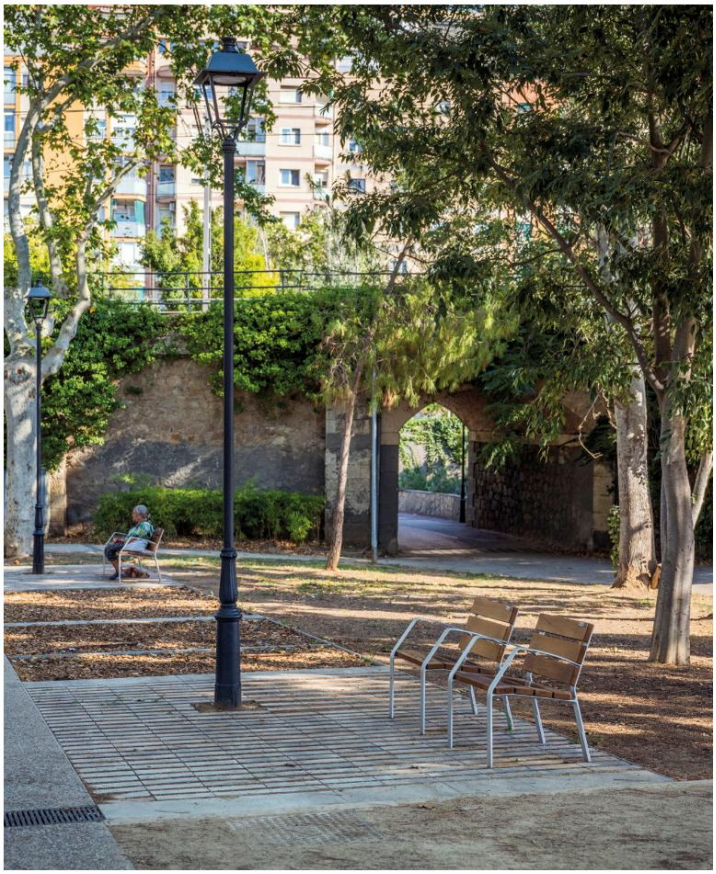
Categoria	Descripció	Exemples representatius	Nre. taules	Estratègia
Dades de domini migrables	Taules que contenen informació històrica del sistema antic	*actions*, *services*, *users*, *fundings*, *projects*, *certifications*, *economic_items*	13	*Script* de migració dedicat
Taules de referència	Diccionaris de valors normalitzats	*action_types*, *cities,* *priorities*, *roles*, *BPM definitions*	17	Còpia directa des de BD de referència
Sense correspondència	Conceptes que no existien al sistema antic	*action_user*, *role_user*, *sub_tasks*, *contacts*, *files*	34	Es poblaran amb l'ús de l'aplicació
Infraestructura del sistema	Taules tècniques del *framework* i motor de regles	*migrations*, *jobs*, *sessions*, *rule_sets*	25	No es migren mai

Taula A.1. Classificació de les 89 taules del sistema destí segons la seva estratègia de migració.

A3. Exemples d'Informes Institucionals

En aquest apèndix es presenten exemples visuals dels informes generats de forma automatitzada pel sistema. S'inclouen tant els gràfics per a presentacions executives (PPTX) com les exportacions detallades (HTML) dissenyades per a una anàlisi més exhaustiva per part dels gestors.

En la Figura A.2 es presenta un exemple d'informe HTML d'obra generat automàticament pel sistema per a l'actuació Millora Parc Can Buxeres (CAS 655). S'hi poden observar a la part dreta les seccions d'estat, finançament, dades tècniques i planificació, amb la fotografia (part esquerra) obtinguda automàticament del servidor d'imatges corporatiu (BIAM).



AMB

Millora Parc Can Buxeres CAS 655

Estat		
Estat [1]	2.6 Retorn garantia	
Estat [2]	Sense dades	
Informació interna AMB: 1 - Estat gestor / 2 - Estat planificació		Act: Sense dades
Finançament OBRA		
Programa finançament AMB	ALTRES AMB PACTE 20-23	185.940 € 900.000 €
Fonts finançament	-	
Dades tècniques		
Servei AMB	SAEP1	
Autor projecte	Equip intern AMB	
Direcció obra	Equip intern AMB	
Constructora	SERVEIS INTEGRALS DE MANTENIMENT RUBATEC, S.A.	
Planificació		
Data signatura contracte redacció	Redacció equip intern AMB	
Data lliurament definitiu projecte execució	16-06-2021 (Real)	
Data aprovació inicial	05-04-2022 (Real)	
Data aprovació definitiva	26-07-2022 (Real)	
Data signatura contracte d'obres	07-11-2022 (Real)	
Data acta de replanteig	16-12-2022 (Real)	
Data acta recepció	18-04-2024 (Real)	
Comentaris		
Sense comentaris		

Figura A.2. Informe HTML d'obra amb fotografia del servidor BIAM (Parc Can Buxeres, CAS 655).

En la Figura A.3 es mostra un exemple de gràfics generats automàticament per a la presentació executiva del programa AMB PACTE 2020-2023. El gràfic esquerre mostra la distribució del nombre d'actuacions per estat i tipologia d'execució; el gràfic dret, la distribució de l'import total d'inversió en les mateixes categories. En tots dos casos s'utilitza un gràfic circular, ja que l'objectiu és destacar una única proporció agregada —el percentatge global d'actuacions executades— de manera immediata i visual. La paleta de color en tons vermells respon a la identitat corporativa de l'AMB.
En canvi, quan cal comparar l'estat d'execució entre un nombre elevat d'entitats simultàniament, el sistema opta per gràfics de barres apilades en lloc de circulars, ja que aquests permeten una lectura comparativa més clara. La Figura A.4 n'és un exemple, aplicat a l'informe d'execució del programa PSA (Pla de Sostenibilitat Ambiental): hi mostra l'estat d'execució de les actuacions per a fins a 20 municipis simultàniament, amb un indicador del percentatge d'obra executada per a cadascun. En aquest cas, la paleta de color en tons verds respon a la identitat temàtica del programa de sostenibilitat, diferenciant-lo visualment dels informes vinculats al programa AMB PACTE.

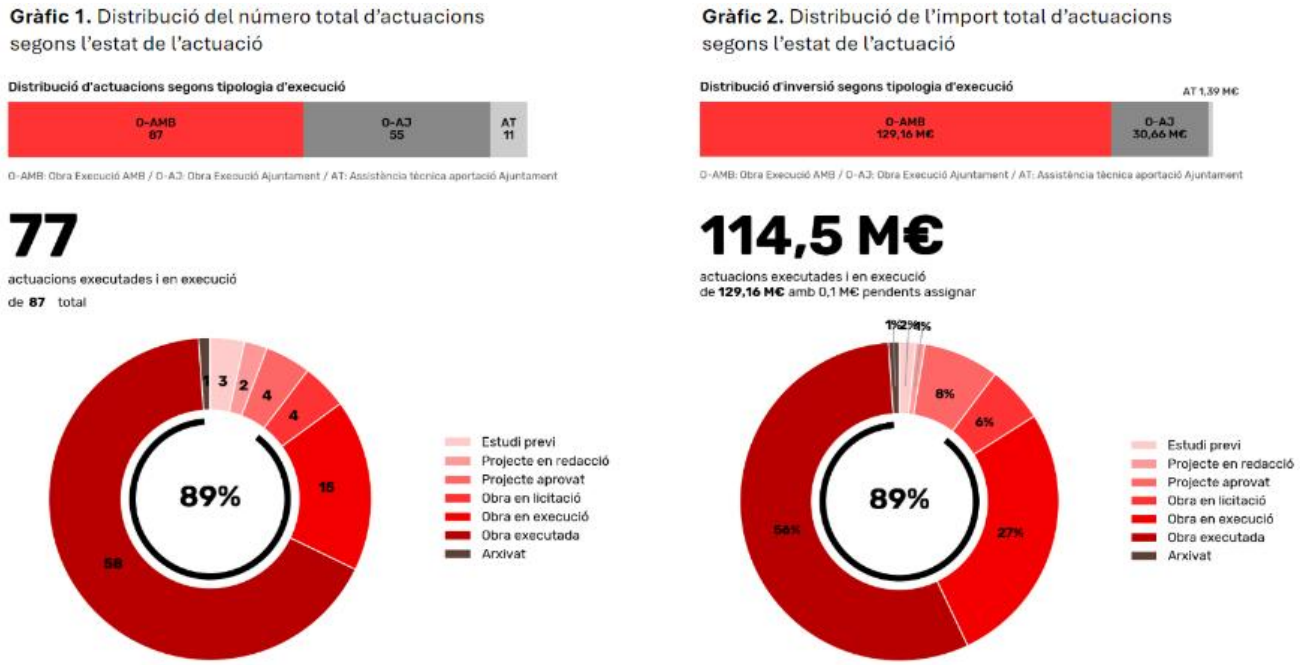


Figura A.3. Gràfics automatitzats de distribució del número d'actuacions i import total d'inversió per al programa AMB PACTE 2020-2023.

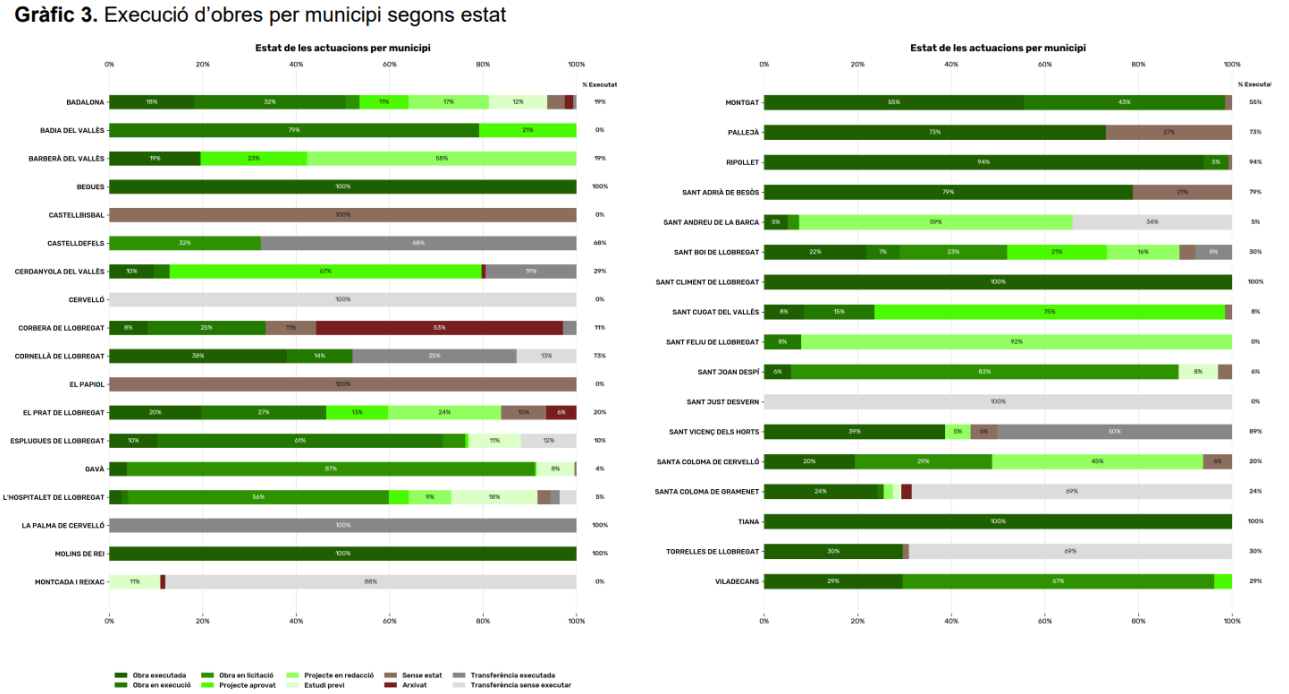


Figura A.4. Gràfic de barres apilades de l'estat d'execució de les actuacions del programa PSA (Pla de Sostenibilitat Ambiental) per municipi, amb indicador del percentatge d'obra executada.